With the support of the Walloon Region of Belgium,
Direction Générale de l'Agriculture

citius

May 2008

A program to apply Markov chain Monte Carlo method for multilocus analysis of large complex pedigrees.

Author: Maciej Szydłowski

E-mail: szydlowski.m@fsagx.ac.be

Introduction

Aim. Citius was developed to solve some genetic problems on large complex pedigrees with incomplete multilocus genotype data. Citius calculates:

- multilocus genotype probabilities for members of a large complex pedigree by the use of the Markov chain Monte Carlo (MCMC) method
- one-locus genotype probabilities for members of a huge complex pedigree by the use of the iterative peeling (IP) approach
- probability of sharing genes identical by descent (IBD) between members of a large complex pedigree at many locations on a chromosome
- Z statistics for the ARP (affected relative pair) analysis based on IBD sharing at multiple points on a chromosome
- parameters of mixed major gene-polygenic multitrait models (gene effects, allele frequency, polygenic additive effects, polygenic additive (co)variance, residual (co)variance).

Citius supports autosomal codominant markers of up to 31 alleles. A multilocus analysis requires that the relative positions of markers on a chromosome are known (known linkage map).

Algorithms. Some of the algorithms used by citius are:

- SPIP genotype sampler (Szydlowski and Gengler 2008)
- genotype elimination (Lange and Goradia 1987)
- allele set-recoding (O'Connell and Weeks 1995)
- iterative peeling (Van Arendonk et al. 1989; Fernandez et al. 2001)
- Metropolis –Hastings independence sampler (Tierney 1994)
- whole meiosis sampler (M-sampler) (Thompson and Heath 1999)
- two- and three-generation family segregation indicator samplers (Thomas et al. 2000) (here the samplers are called A2 and A3).
- random number generators from RANLIB.C - Library of C Routines for Random Number Generation (http://www.stat.umn.edu/HELP/ranlib-docs/ranlib.c.chs) and from Matsumoto and Nishimura (1998).

Samplers. For genotypes citius uses 6 basic samplers:

- L-sampler. This sampler updates genotype configuration at a single locus for entire pedigree. The sampler does not require an initial genotype configuration of genotypes and it is always used first at start. The sampler is irreducible. The

sampler draws a sample from approximate distribution (inexact sampler). The approximate probabilities are calculated by the alternate use of two algorithms: simple peeling and iterative peelingas discribed by Szydlowski and Gengler (2008). The simple peeling algorithm calculates exact probabilities for unlooped part of a pedigree. Further, the calcuation is continued by the use of the iterative peeling algorithm. This way the inexact probabilities are calculated. A single individual from a pedigree loop is selected and its genotype is sampled from the inexact probability. The sampled genotype 'breaks' the loop and calculation is continued by the use of the simple peeling algorithm until next loop is encountered. The process continues until all loops are broken and all probabilities are calculated. The random sample is taken conditionally on already sampled genotypes. The sampler is designed for large complex pedigrees, for which calculation of exact probabilities are difficult. For multilocus problems, citius uses the L-sampler for each locus in turn. For linked markers, however, the sampler mixes badly and the multilocus samplers (M, A2,and A3) must be used to improve mixing.

- S-sampler. This is single site sampler. The sampler draws random genotype for an individual at a single locus. Whole pedigree is updated at locus by visiting all individuals in turn. It may be reducible and it does not mix well in large pedigrees, even in case of single marker problem. In general, this sampler is not recommended.
- M-sampler. This sampler was proposed by Thompson and Heath (1999). The sampler updates segregation indicators at whole meiosis. It is designed to improve mixing in multilocus problems. The sampler is reducible and should always be used in a combination with the L-sampler.
- A2-sampler. This sampler was proposed by Thomas et al. (1999). The sampler updates segregation indicators in two-generation families (being a part of a large pedigree). The sampler is irreducible and should always be used in a combination with the L-sampler. The sampler can be used to improve mixing in multilocus problems.
- A3-sampler. This sampler was proposed by Thomas et al. (1999). The sampler updates segregation indicators in three-generation families (being a part of a large pedigree). The sampler is reducible and should always be used in a combination with the L-sampler. The sampler can be used to improve mixing in multilocus problems.
- For polygenic additive effects citius uses a sampler, which for each individual samples random effect across all traits at once.

Input. Citius uses the following text files for input:

- control file (commands for citius) (required)
- pedigree file (required)

- genotype data set (optional)
- phenotypes (quantitative traits) (optional)
- 'affected' individuals (optional)

Output. The results are sent to text files. The following text files are produced depending on the analysis:

- 'out.arp' - results from IBD analysis calculated by the use of the MCMC method-approximate Z statistics for ARP (affected relative pair) analysis
- 'out.fix' - values of fixed effects
- 'out.freq' - values of allelic frequencies
- 'out.fset' - the set of feasible genotypes for all pedigee members at all loci, as deduced by the use of genotype elimination algorithm
- 'out.gp' - estimates for genotype probabilities calculated by the use of the MCMC method. This file may contain multilocus genotype probabilities.
- 'out.ibd' - results from IBD analysis calculated by the use of the MCMC method - IBD sharing estimates
- 'out.ip' - estimates of genotype probabilities calculated by the use of the iterative peeling method. This file may contain genotype probabilities for many loci, however, the loci are assumed unlinked.
- 'out.log' - messages, warnings and errors
- 'out.padd' - estimates of polygenic additive effects for individuals
- 'out.qtg' - values of homozygous and heterozygous effects of genes
- 'out.va' - values of polygenic additive (co)variance matrix
- 'out.ve' - values of residual (co)variance matrix

Installation. The program is written in C language and was tested on a Pentium microcomputer running Linux. The gcc compiler was used. There is no gurantee that the program can be compiled and run under other systems.

(1) use **tar zxf citius.tar.gz** to untar the archive file

(2) use **make** to compile the program - executable file **citius** is created

Running. To run the program use **./citius** *control_file*. You may want to run the program with some other seed for the random number generation that the default seed. In this case type **./citius -s** *seed control_file*, where the *seed* is some arbitrary string. The verbosity level can be controlled by **-v** option. Total silence (no output sent to screen) is reached with **-v** *0* option.

Errors. If you find an error in citius, I would appreciate an e-mail message from you.

Please send the message to mcszyd@jay.au.poznan.pl

Citation. If you use citius in your scientific work for MCMC genotype sampling, please cite the following paper:
Szydlowski M, Gengler N (2008) Sampling genotype configurations in a large complex pedigree. Journal of Animal Breeding and Genetics (accepted).

Pedigree

Preparing pedigree file. Pedigree information is in a file separate from genotype records. The file is in free format, i.e., all variables are separated by spaces. The pedigree file includes at least three fields: an individual, its father and mother. If multiple genetic groups are considered, one additional field has to identify the genetic group for the individual. An individual's code (ID) is a string of maximum length of 20. The ID must be unique for the entire pedigree, even if there are multiple independent families in the pedigree. A missing parent or missing genetic group is coded as 0 (zero). The order of all fields is arbitrary but should be the same for all records. By default citius expects the following order: (1) individual, (2) father, (3) mother. If the order in the pedigree file is different or the additional field for genetic group is also included the PEDIGREE_INPUTS command should be used in the control file to instruct citius how to read the file. The pedigree file is introduced to citius by the use of the PEDIGREE_FILE command.

Missing parents. A missing parent has the code 0 (zero). In general, an individual with progeny but both parents missing does not have to have its own record. If multiple genetic groups are considered, however, such an individual should also be included with a field for its genetic group. If both parents of an individual are unknown, that individual is considered a founder. A founder with no progeny has no ties, such an individual should be removed from pedigree file.

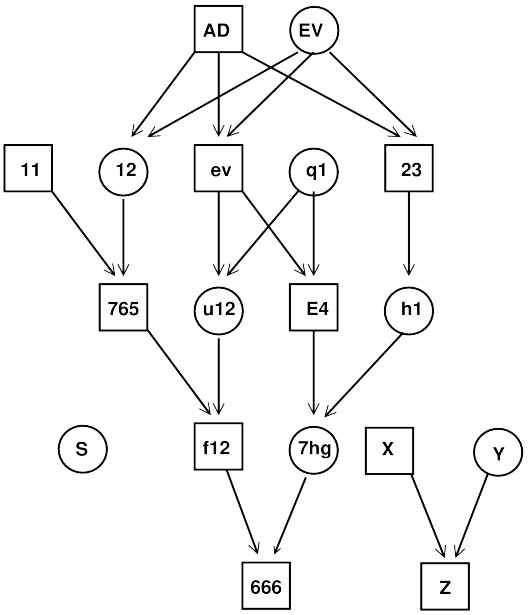
Artificial parent. If only father or only mother is missing for a child, an artificial parent is automatically created. The artificial parents are created by citius to make computation simpler. The artificial parents may appear in output files. The code for the dummy father is the child's code+_F, and for the artificial mother is the child's code +_M. For example, if an individual h1 has known father and missing mother, the code for the missing mother is h1_M.

Multiple genetic groups. Multiple genetic groups can be considered if founders come from different populations or come from the same population but were born in different years and the population underwent significant genetic changes. If multiple genetic groups are used it is possible to assume that an allele occurs with different frequency in various genetic groups.

If only a single genetic group is assumed, the group does not have to appear in a pedigree file - all founders are assigned automatically to single default genetic group and the group name is simply 'default'. If multiple genetic groups are assumed, however, the information on genetic group for a founder must appear in a pedigree file. A founder assigned to some genetic group must have its own record in a pedigree file and a name of genetic group must appear within its record. All genetic groups must have its own unique codes. A code is a string of maximum length of 20.

If you assume multiple genetic groups you should include the PEDIGREE_INPUTS command in the control file. The command specifies the column occupied by genetic

Example. The pedigree and the corresponding pedigree file. The population consists of 2 families and an individual with no ties (S). The individual S has no ties and therefore it is not included in the pedigree file. Two seperate families are included. These families are independent and you may want to analyze them separately by creating two pedigree files. In this example, however, they are analyzed together. Families have no name or number, therefore, it is important to have a unique code for each individual. A code cannot be used for more than one individual even if the individuals belong to seperate families. Note, the individuals EV and ev are two different members of the same family. The female h1 has only her father known and her mother is missing - the missing mother is coded as 0 (zero). The order of records within a pedigree file does not matter. All founders (fathers and mothers with their own parents unkown) are assigned to a default genetic group. The pedigree file is 'e1.ped'. In the control file (only a part of the control file is shown), the command PEDIGREE_FILE is used to introduce the pedigree file to Citius.




```
12 AD EV
ev AD EV
23 AD EV
765 11 12
u12 ev q1
E4 ev q1
h1 23 0
f12 765 u12
7hg E4 h1
666 f12 7hg
Z X Y
```

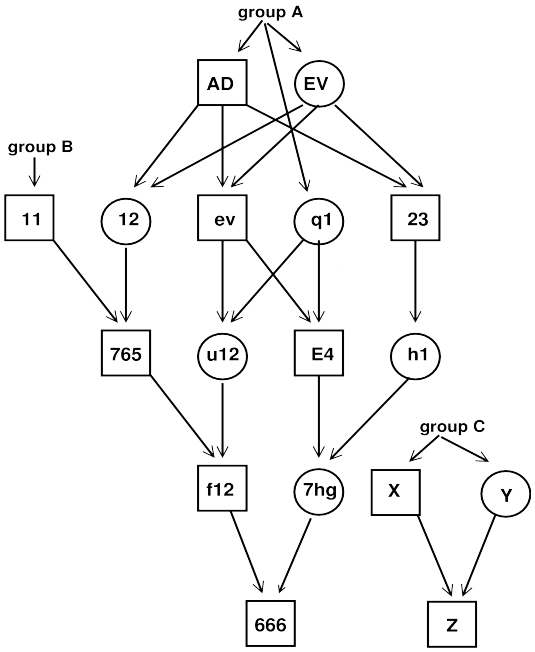
```
pedigree_file  e1.ped
...
```

group and other pieces of information. The default sequence is: IFM (individual, father and mother), and it does not include the genetic group, and consequently multiple genetic groups are not used. The following command

PEDIGREE_INPUTS IFMG

overwrites the default sequence and specifies that the genetic group (G) is in the fourth column. If a founder has its genetic group unknown (code 0), it is automatically assigned to a default genetic group. The name of the default genetic group is simply 'default'.

Example. Multiple genetic groups. The founders were assigned to three different genetic groups: A, B and C. The founders must now have their own records in the pedigree file. If a record for nonfounder also includes the field for its genetic group, the field is skipped. In the control file (only a part of the file is shown) the PEDIGREE_INPUTS command indicates that the genetic group occupies the fourth field.




```
AD 0 0 A
EV 0 0 A
11 0 0 B
12 AD EV
ev AD EV
23 AD EV
765 11 12
q1 0 0 A
u12 ev q1
E4 ev q1
h1 23 0
f12 765 u12
7hg E4 h1
666 f12 7hg
Z X Y
X 0 0 C
Y 0 0 C
```

```
pedigree_file  e1.ped
pedigree_inputs IFMG
...
```

Genotype records

Citius uses autosomal codominant markers. Markers on X and Y chromosomes and dominant markers are not supported. Up to 31 alleles can be used. For a multilocus analysis the order and positions of markers on genetic map must be known.

Preparing data file. Genotype data is in a file separate from pedigree information. The file is in free format, i.e., all variables are separated by spaces. The data file includes the following fields: ID, allele 1 at marker 1, allele 2 at marker 1, allele 1 at marker 2, allele 2 at marker 2, etc. An ID is a string of maximum length of 20, the code is the same code as used in accompanying pedigree file. The ID cannot be coded as missing (zero). A record for an individual that is not included in the accompanying pedigree file is simply ignored. An allele is coded with a positive number (integer). A missing allele must be coded as 0 (zero). If one allele is missing the other allele is ignored. Other words, an individual with one allele missing is treated as an unobserved individual at a locus.

You should specify the path to the genotype data file in the control file for citius. For this, use the GENOTYPE_DATA_FILE command. You should also specify how many markers should be read from the data file. For this, use the NUMBER_OF_MARKERS command followed by the number of loci. Markers are read from the file according to the POSITIONS_IN_FILE command.

Order of markers. The order of loci in a genotype file does not have to be the true order on a chromosome. You can specify the true genetic order by providing the position of each marker on the genetic map. For this use the MARKER_SPEC command. Multiple MARKER_SPEC commands can be used - one command for each marker to be considered. The MARKER_SPEC command has several options (subcommands): option NAME is used to give a name to a marker, the option POSITION is to describe the positions (in centiMorgans) of the marker on male and female maps. In the example the MARKER_SPEC command appears three time because three markers are used. First marker in the genotype data file 'e1.dat' is D8S552, second is D8S351 and the third is D8S261. The positions of the markers on male and female genetic maps are provided. Given the positions, the order of markers is D8S351-D8S552-D8S261. Note, the order of markers as concluded from the given positions of markers on the male map must be the same order as can be concluded from the positions on the female map.

Example. The example genotype records and partial control file. The data consists of two records for two individuals u12 and 666. Genotypes at three polimorphic sites are given (columns 2-7). The alleles are coded with arbitrary positive integer numbers. In the accompanying control file the name of the data file is specified by the GENOTYPE_DATA_FILE command and the number of markers to be read from the data file is specified by the NUMBER_OF_MARKERS command. The names and positions of genetic markers are specified by the use of the MARKER_SPEC command. The MARKER_SPEC command is used four times. Within each MARKER_SPEC command, the name of marker follows the NAME command, the position of the marker on male and female map follow POSITION command, and the fields in the file occupied by a marker data follow the POSITIONS_IN_FILE command. Two zeros following the POSITIONS_IN_FILE command mean that a marker is not observed (useful for modelling a hypothetical gene for quantitative trait).

```
u12  12 11  123 124  2 3
666  0 12   123 124  5 2
```

```
....
pedigree_file  e1.ped
genotype_data_file  e1.dat

number_of_markers  4

marker_spec
   name D8S552
   position 19.5 28.4
   positions_in_file  2 3
marker_spec
   name D8S261
   position 10.4 16.8
   positions_in_file  5 6
marker_spec
   name D8S351
   position 25.6 37.6
   positions_in_file  4 5
marker_spec
   name Hypothetic
   position 29.0 40.0
   positions_in_file  0 0
....
```

Phenotypic records

Quantitative trait records can be analyzed in:
- multitrait polygenic additive models
- multitrait monogenic, oligogenic or pure polygenic models
- multitrait mixed major genes-polygenes inheritance models

Homozygous and heterozygous effects of a gene can be assigned to an observed marker or a hypothetical gene.

Citius can estimate the following
- polygenic additive (co)variance matrix
- residual (co)variance matrix
- polygenic additive individual effects
- effects of genotypes at observed or hypothetical markers

If a locus is labelled Y_FUNCTIONAL, phenotypic records are considered during genotype probability calculations and genotype sampling. If a marker is assumed to be a functional polymorhism effecting a quantitative traits (phenotypes), the trait records are taken into account when genotype probabilities are calculated and missing genotypes are sampled. In this case, the probability for missing genotype is estimated based on observed genotypes of relatives and the observed quantitative trait. Because many calculations are based on genotype samples or genotype probabilities, including IBD analyses, such quantitative trait may shape final results produced by citius. The iterative peeling algorithm (used for calculation of genotype probabilities without MCMC sampling) also takes into account availabe quantitative records. It is assumed that within a particular genotype, a trait follows normal distribution.

Preparing data file. Phenotype data is in a file separate from pedigree information. The file is in free format, i.e., all variables are separated by spaces. The data file often includes the following fields: ID (individual), fixed effects and recorded traits. An ID is a string of maximum length of 20, the code is the same code as used in accompanying pedigree file. The ID cannot be coded as missing (zero). A record for an individual that is not included in the accompanying pedigree file is simply ignored. You should specify the path to the phenotype data file in the control file for citius. For this, use the PHENOTYPE_DATA_FILE command.

Reading data file. You should instruct citius how to read your data file. To instruct citius, you should use the PHENOTYPE_DATA_INPUTS command followed by your instruction. The instruction is a string of the following letters:
I - individual (ID)
Y - quantitative trait to be analysed
F - classification variable for a fixed effect
S - unimportant variable to be skipped

The order of letters corresponds to columns in your data file.

Model for a quantiative trait. When records for quantitative traits are available, the default model is a pure polygenic additive model. If an observed or hypothetical marker is assumed to influence a trait, you should use for the marker the command Y_FUNCTIONAL. There can be one, two or more functional genes for a trait. To remove a polygenic effect from a model you can use the command 'POLYGENIC_EFFECT_MODE frozen'. The polygenic effects are frozen in their intitail values, which are all zeros.

Example. The example phenotype records and partial control file. The file contains 6 columns (fields): (1) individual's name, (2) trait, (3) classification variable, (4) trait, (5) trait, (6) other variable. In a control file, you should include the PHENOTYPE_DATA_FILE command followed by the file name. The PHENOTYPE_DATA_INPUTS command determines which traits and classification variables should be read and included in the statistical model. A user specified ISFYY. This means that individual's name can be found in the first column (I). The second column should be skipped (S). The third column contains an important classification variable for fixed effect (F). The next two columns are quantitative traits (YY). The last column in the file is ignored (for clarity a user may type ISFYYS). The missing observation is coded as '000'. Initial (co)variance matrices for polygenic additive and residual effects are provided (lower triangle only!). Therefore, the statistical model is a two-trait model with two fixed effects (including general mean and season), polygenic additive effect (Animal Model) and a direct effect of RYR1 marker. If all genotypes at RYR1 locus were known, they could be included as fixed effect in the phenotype data file. Here, some genotypes were missing (or one wants to differentiate between CT and TC heterozygotes) and had to be sampled. The observed genotypes were included in a genotype data file (not shown).

```
 u12  16.0  winter  0.1 220 4
 666  18.69 summer  0.5 221 4
 AD   28.5  summer  0.3 000 3
 7hg  22    spring  0.3 224 4
....
```

```
...
phenotype_data_file  /mydir/mytraits.txt
phenotype_data_inputs ISFYY
missing_y_code 000

polygenic_variance
0.01
0.0 10

residual_variance
0.01
0.0 22

number_of_markers 1
marker_spec name RYR1 y_functional
```

Affecteds

A file being a list of the individuals of interest (affecteds) can be provided by the user. The file is optional but helps to save computer memory and reduce time of an analysis. You can use the file if output statistics are required for a part of the pedigree rather than for entire population. If the file is provided the entire pedigree is analyzed but the output statistics are considered only for ‘affected’ individuals. The file of 'affecteds' is especially important for IBD analysis in a large pedigree. By default, IBD sharing statistics are calculated for all possible pairs of pedigree members. The number of different pairs (related and unrelated) can be huge. If the file, beilng a list of 'affecteds' is introduced to citius, the program analyzes only pairs of affected individuals. If you want to use a file with the list of 'affected' individuals it is important to introduce its name in the control file for citius. For this, the AFFECTED_INDIVIDUALS_FILE command is used.

Preparing the file. The file is a simple list of individuals of interest prepared as a text file. All IDs can be separated by space or new line sign. An ID included in the file must also appear in the accompanying pedigree file. An individual that appear in the file but is not included in the pedigree file is ignored.

Example. A partial control file. The AFFECTED_INDIVIDUALS_FILE command is used to specify the name of the file 'e1.aff' being the list of individuals of interest.

```
....
pedigree_file e1.ped
genotype_data_file  e1.dat
affected_individuals_file e1.aff
....
```

Control file

A control file consists of commands for controlling the run of citius. The following commands can be used:

Input pedigree
- PEDIGREE_FILE …
- PEDIGREE_INPUTS …

Input markers
- GENOTYPE_DATA_FILE …
- NUMBER_OF_MARKERS …
- MARKER_SPEC
--- POSITIONS_IN_FILE …

Input affecteds
- AFFECTED_INDIVIDUALS_FILE …

Input phenotypes
- PHENOTYPE_DATA_FILE …
- PHENOTYPE_DATA_INPUTS …
- MISSING_Y_CODE …

Input model parameters
- NUMBER_OF_MARKERS
- MARKER_SPEC
--- POSITION …
--- FIX_ALLELE_FREQ …
--- GENE_EFFECTS …
- ADDITIVE_VARIANCE …
- RESIDUAL_VARIANCE …

Model for quantitative trait(s)
- PHENOTYPE_DATA_INPUTS …
- MARKER_SPEC
--- Y_FUNCTIONAL
--- NO_IMPRINTING
- POLYGENIC_EFFECTS_MODE
- ADDITIVE_VARIANCE …
- RESIDUAL_VARIANCE …
- USE_MATERNAL_ADDITIVE_EFFECT

IBD analysis
- IBD_POINTS …
- IBD_WALK …
- IBD_GAMETIC_MATRIX

MCMC multilocus genotype probability
- MARKER_SPEC
--- G_PROB

IP genotype probability
- MARKER_SPEC
--- ITERATIVE_PEELING

Markov chain
- BURN_IN …
- NUMBER_OF_SAMPLES …
- THINNING_VALUE …
- LSAMPLER_USE …
- SSAMPLER_USE …
- MSAMPLER_USE …
- A2_USE …
- A3_USE …

Parameter updates
- FIX_EFFECT_MODE …
- GENE_EFFECT_MODE …
- POLYGENIC_EFFECT_MODE …
- POLYGENIC_VARIANCE_MODE …
- RESIDUAL_VARIANCE_MODE …
- MARKER_SPEC
--- ALLELE_FREQ_MODE
- IP_MAXIMUM_ITERATIONS …
- POLYGENIC_LOOP_ITERATIONS …

Output to screen
- DISPLAY_VALUE …

Output to files
- PRINT_FIX_EFFECTS
- PRINT_GENE_EFFECTS
- PRINT_POLYGENIC_VARIANCE
- PRINT_RESIDUAL_VARIANCE
- ESTIMATE_POLYGENIC_EFFECTS
- IBD_POINTS …
- IBD_WALK …
- IBD_GAMETIC_MATRIX

- MARKER_SPEC …
--- PRINT_ALLELE_FREQ
--- G_PROB
--- ITERATIVE_PEELING

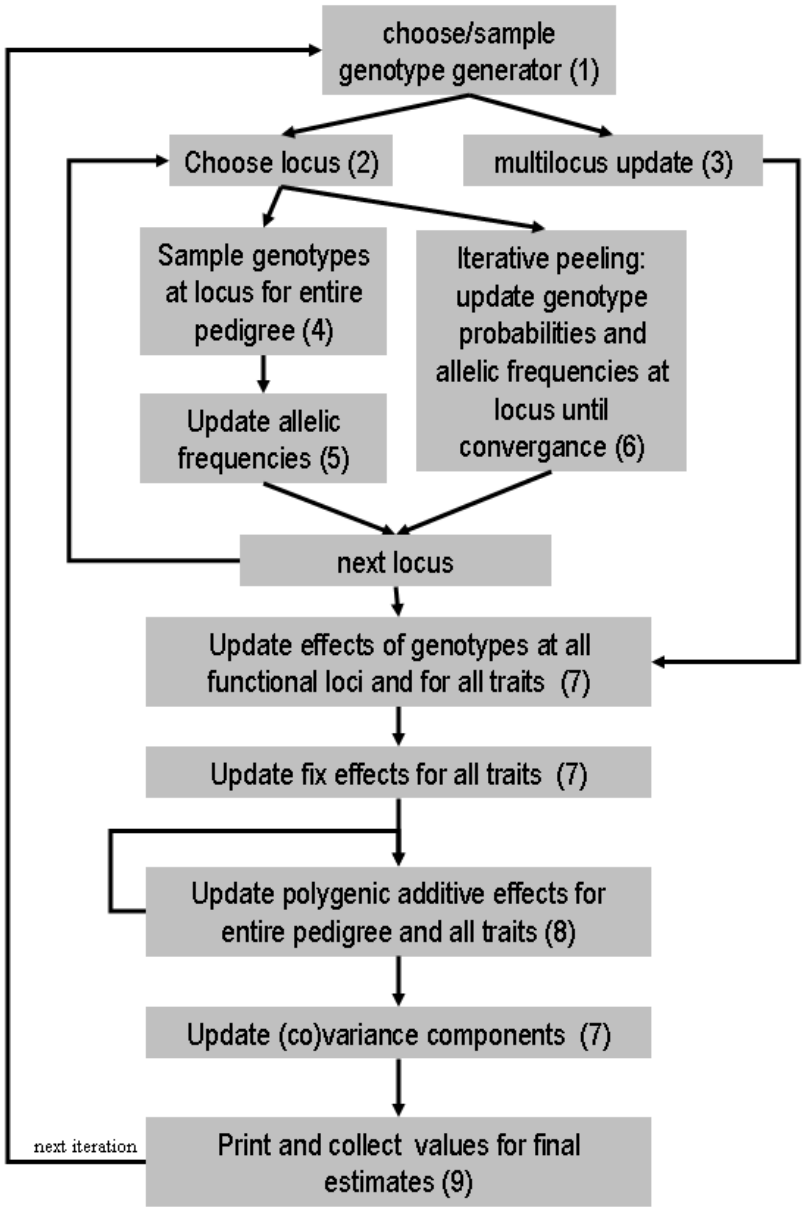


Figure. Simplified algorithm of main loop. (1) For multilocus MCMC samplers are chosen randomly. (2) Loci are updated in random order. (3) Three different multilocus genotype updates are possible: A2, A3 and M. (4) Joint configuration of genotypes at locus is generated via L sampler. From second iteration on also S sampler can be used. (5) Some or all allele frequencies can be fixed at know values. (6) For non MCMC analysis. (7) When incremental mode is used, the generator updates values every (n*k) iteration based on average of n expected values collected every k-th iteration. You can set n by using the NUMBER_OF_INCREMENTS command. (8) You can set the number of internal iterations for polygenic effects by using POLYGENIC_LOOP_ITERATIONS command. (9) This step is performed every k-th iteration after burn-in period. Set burn-in period using BURN_IN command. Set k by using THINNING_VALUE command.

Commands in alphabetical order

A2_USE specifies the proportion of the use of the A2-sampler for missing genotypes. The command is used in combination with the LSAMPLER_USE, SSAMPLER_USE, MSAMPLER_USE and A3_USE commands. For example, the following combination of commands: LSAMPLER_USE 10 MSAMPLER_USE 3 A2_USE 1 A3_USE 0 specifies the proportion of L-, M-, A2- and A3-samplers to be 10:3:1:0. Therefore, at an iteration, the probability to use the A2-sampler is 1/14**.**

A3_USE specifies the proportion of the use of the A3-sampler for missing genotypes. The command is used in combination with the LSAMPLER_USE, SSAMPLER_USE, MSAMPLER_USE and A2_USE commands. For example, the following combination of commands: LSAMPLER_USE 10 MSAMPLER_USE 3 A2_USE 2 A3_USE 1 specifies the proportion of the L-, M-, A2- and A3-samplers to be 10:3:2:1. Therefore, at an iteration, the probability to use the A3 sampler is 1/16**.**

ADDITIVE_VARIANCE command serves to input polygenic additive (co)variance matrix for quantitative traits analyzed under Animal Model. The number of traits is decided by using the command PHENOTYPE_DATA_INPUTS. Only lower trangle of a matrix should by given. The matrix can be used as a initial value or kept unchanged (known variance) during calculations (see polygenic_VARIANCE_MODE). Beside a matrix for direct (D) additive effects it is also possible to input maternal (M) additive (co)variance matrix and covariance between direct and maternal effects (DM). In this case the entire matrix is:

D (MD)
(DM) M

Again, only lower triangle of the matrix should be provided.

AFFECTED_INDIVIDUALS_FILE specifies the optional file being a list of the individuals of interest. If the file is provided, the output results are printed only for 'affected' individuals rather than the entire pedigree. For a very large pedigree the file of 'affecteds' can save time and memory.

ALLELE_FREQ_MODE is a subcommand used within the MARKER_SPEC command. See the MARKER_SPEC command for details.

BURN_IN specifies the number of samples for burn-in period. Samples in burn-in period are ignored. Note, citius does not check MCMC for convergence. For unilocus analysis under fixed allele frequencies, citius needs no burn-in period because genotype samples are sampled from desired probability distribution. If the allelic frequencies remains unknown, citius needs some burn-in period. The length of burn-in period depends on particular pedigree and data. Note, there is no reliable method to prove convergence of MCMC method. The possible option is to print samples of gene frequency and monitor stationarity of the sampling process. During burn-in period, no parameter values are printed to files.

DISPLAY_VALUE specifies how often citius sends output to screen. The program displays information on the current number of iteration and samples/values collected/printed. For example, DISPLAY_VALUE 5 specifies that screen is updated every 5th iteration. Note, for total silence use –v 0 argument at the command line.

FIX_ALLELE_FREQ is a subcommand used within the MARKER_SPEC command. See the MARKER_SPEC command for details.

GENE_EFFECTS is a subcommand of the MARKER_SPEC command. See the MARKER_SPEC command for details.

G_PROB is a subcommand of the MARKER_SPEC command. See the MARKER_SPEC command for details.

GENOTYPE_DATA_FILE Specifies the file with genotype records. The command should be followed by the path to the data file.

IBD_GAMETIC_MATRIX specifies that IBD relationship is computed between gametes rather than individuals. The command is used together with IBD_POINTS or IBD_WALK commands.

IBD_POINTS performs IBD analysis at selected points on a chromosome. The ibd_points command is followed by the required number of points and positions on male and female linkage map (in cM). For example, the following command ibd_points 3 5 8 12 20 17 25 means that IBD sharing will be calculated at three points. On the male map the points are 5, 12 and 17 cM, and on the female map these are 8, 20 and 25 cM. Certainly, a point on the male and female map should be flanked by the same pair of markers. The command is alternative to IBD_WALK.

IBD_WALK performs IBD analysis for evenly spaced points on a chromosome. It specifies the starting point, the step, and the final point on a chromosome. For example, the following command ibd_walk 5 3 17 means that the points to be analyzed are: 5, 8, 11, 14 and 17 (on sex average map). The command is alternative to IBD_POINTS.

IP_MAXIMUM_ITERATIONS specifies maximum number of iterations for the iterative peeling algorithm. If this command is not included in the control file, citius iterates until convergence or until some default maximum value of iterations is reached. The specified value overwrites the default one. Note, citius uses the iterative peeling algorithm for the MCMC method to calculate approximate genotype probability distribution to draw random genotype configuration. In case of the MCMC method, manipulating with maximum number of iterations may ruin sampler performance.

ITERATIVE_PEELING is a subcommand of the MARKER_SPEC command. See the MARKER_SPEC command for details.

LSAMPLER_USE specifies the proportion of the use of the L-sampler. The command is used in a combination with the MSAMPLER_USE, A2_USE and A3_USE commands. For example, the following combination of commands: LSAMPLER_USE 10 MSAMPLER_USE 3 A2_USE 2 A3_USE 1 specifies the proportion of L-, M-, A2- and A3-samplers to be 10:3:2:1. Therefore, at an iteration, the probability to use the L-sampler is 10/16**.**

MARKER_SPEC command allows using marker specific information and model. The command should be used as many times as it is specified by the NUMBER_OF_MARKERS command. There are 12 (sub)commands within the MARKER_SPEC command:

> **ALLELE_FREQ_MODE** (1 argument) instructs citius how allelic frequencies should be updated. The following options are allowed: *freeze* (or *frozen*) - all allele frequencies are constant, *sample* or *random* (default) - current allele frequency is replaced by a random sample, *expected* - current value is replaced by an expected value calculated from current sampled/observed genotypes or genotype

probabilities (if ITERATIVE_PEELING is used) in base population, *increment* - allele frequency is updated by temporary estimate every *n* iteration. The estimate is calculated as an average of expected values from previous *n* iterations (starting from zero, in *i*th iteration a new estimate of allele frequency (*p*) is incremented by $E_i(p)/n$). See also FIX_ALLELE_FREQ and NUMBER_OF_INCREMENTS subcommands.

FIX_ALLELE_FREQ (>=4 arguments) can be used to fix some allelic frequency at some known values. The command is followed by the number of fixed values. Then, for each value to be fixed, a user should specify the genetic group (as used in pedigree file), the allele code (as used in data file) and finally the allele frequency. If no genetic group was assigned to an individual in a pedigree file, the default genetic group is assaigned automatically. The default genetic group is named 'default'.

GENE_EFFECTS (>=3 arguments) can be used together with Y_FUNCTIONAL. It specifies initial/known gene effects on quantitative trait(s). This subcommand can be used only for biallelic locus (including an unobserved hypothetical gene). For one-trait model, this command should be followed by three effects for three genotypes 0/1, 1/0, 1/1 (orded is important!). The effect of genotype 0/0 is always zero and may not be provided. For a model with n traits, the gene_effects command should be followed by 3×n values (n values for genotype 0/1, n values for genotype 1/0, and n values for genotype 1/1). Here the genotypes are described by using citius internal codes for alleles (0 and 1 for biallelic locus). The code 0 is assigned to an allele with the lowest original code. For example, if original codes are 241 and 243, the internal code 0 is for allele 241. See also GENE_EFFECT_MODE.

G_PROB (no arguments) specifies that a marker is considered in MCMC multilocus genotype probability estimation.

ITERATIVE_PEELING (no arguments) command should be used to calculate genotype probabilities for a marker by using the iterative peeling (IP) approach. When the IP is used for a marker, missing genotypes are not sampled. In consequence, frequency of alleles and gene effects are updated based on current genotype probabilities rather than samples. The final genotype probabilities are printed to file 'out.ip'. The IP method can be used for many markers, however, all markers are assumed independent (unlinked). Therefore, probabilities for multilocus genotypes cannot be calculated with IP approach. The iterative peeling method is not MCMC analysis, however, for a very large complex pedigree the IP method can be the only possibility to calculate approximate genotype probabilities. citius performs IP until convergence or until default maximum number of iterations is reached. Estimation of genotype effect on a trait from individual marginal genotype probabilities is questionable. See also

IP_MAXIMUM_ITERATIONS.

NAME (1 argument) is obligatory and should be followed by an arbitrary name assigned to a marker. A name of a marker is a single string that cannot be longer than 20 characters.

NO_IMPRINTING (no arguments) specifies that two alternative heterozygotes (1/2 and 2/1) have the same genotypic effect on quantitative trait. If this command is not used, citius calulates seperate genotypic effects for alternative hetorozygotes.

NO_EXTRA_ALLELE (no arguments) specifies that no extra allele is considered. Be default citius adds one extra allele to computation. Such extra allele represents all hypothetical alleles which possibly exist but are not observed. The extra allele is coded as 0 (zero). If this command is found in a control file, the extra allele is not used. If it is know that all possible alleles at a locus have been observed, the computational requirements can sometime be reduced including this command in a control file.

NO_OVERDOMINANCE (no arguments) forces genotypic value of a heterozygote to be a value within a range defined by the two homozygotes. This option is useful for an hypothetical gene. Giving too much freedom to parameters of an hypothetical gene may result in incorrect estimates.

NO_SET_RECODING (no arguments) disables allele set-recoding. The set-recoding of alleles reduces computational burden for sparse data and is used by default. Set recoding, however, should be disabled for calculation of genotype probability by using iterative peeling approach (see ITERATIVE_PEELING subcommand) and when a marker is labelled Y_FUNCTIONAL. If genotype probabilities are calculated by the use of the iterative peeling approach (ip_genotype_probability command), the calculated probabilities are also defined on set-recoded alleles. Such results are useless.

POSITION (2 arguments) is obligatory for multilocus analysis. It should be followed by two values being the postition of the marker on male and female linkage map (in centiMorgans). For sex average map use two identical values. If only one marker is used, its position has no meaning.

POSITIONS_IN_FILE (2 arguments) should be followed by two numbers being the columns (fields) in the genotype data file where the two observed (unordered) alleles are stored. Example: positions_in_file 2 3 means that for the current marker the observed alleles should be read from columns 2 and 3. Note, positions_in_file 0 0 means that the current marker is not observed. Unobserved markers can be used if you want to check whether one or more hypothetical loci shape some observed phenotype (quantitative trait).

PRINT_ALLELE_FREQ (no arguments) forces citius to print allelic frequencies of a marker to a file. The output is sent to 'out.freq' file.

Y_FUNCTIONAL (no arguments) specifies that a marker or an unobserved hypothetical gene shapes the observed quantitative trait(s). This command is useful for association study between quantitative trait and a particular gene polymorhism when some genotypes are missing. Genotypic effects for a marker can be estimated. If a marker really influences the observed traits, the collected phenotypic records may add extra information on possible genotype of an untyped individual. It is assumed that, within a genotype, traits follow multivariate normal distribution. In segregation analysis, when an unobserved hypothetical gene is labelled Y_FUNCTIONAL, the gene is assumed biallelic (4 genotypes possible). See also NO_SET_RECODING.

MSAMPLER_USE specifies the proportion of the use of the M-sampler. The command is used in combination with the LSAMPLER_USE, SSAMPLER_USE, A2_USE and A3_USE commands. For example, the following combination of commands: LSAMPLER_USE 10, Ssampler_use 0, MSAMPLER_USE 3 A2_USE 2 A3_USE 1 specifies the proportion of L-, S-, M-, A2- and A3-samplers to be 10:0:3:2:1. Therefore, at an iteration, the probability to use the M-sampler is 3/16**.**

NAME is a subcommand used within the MARKER_SPEC command. See the MARKER_SPEC command for details.

NO_EXTRA_ALLELE is a subcommand used within the MARKER_SPEC command. See the MARKER_SPEC command for details.

NO_OVERDOMINANCE is a subcommand used within the MARKER_SPEC command. See the MARKER_SPEC command for details.

NO_IMPRINTING is a subcommand used within the MARKER_SPEC command. See the MARKER_SPEC command for details.

NO_SET_RECODING is a subcommand used within the MARKER_SPEC command. See the MARKER_SPEC command for details.

NUMBER_OF_INCREMENTS (1 argument) specifies the number of values used to update a paremeter by its expectation. Expected value is calculated as a mean from some number of calculated values for the parameter.

NUMBER_OF_MARKERS (1 argument) specifies the number of markers to be read from a data file and then used in analysis. Markers are read from the left to the right. If there are more markers in the data file than the specified number, the additional markers are ignored.

NUMBER_OF_SAMPLES specifies the total number of samples to be used for MCMC estimation. Note, the samples within burn-in period and the samples not accepted in Metropolis-Hastings step are not used.

PEDIGREE_FILE (1 argument) specifies the pedigree file name. The command should be followed by path to the pedigree file.

PEDIGREE_INPUTS (1 argument) describes pedigree file structure. The command should be followed by a string constructed from the following letters: I (individual), F (father), M (mother), G (genetic group, optional) and S (skip, optional). The string IFM is a default one and means that the first three fields (columns) in a pedigree file are codes for an individual, father and mother. If you want to consider genetic groups you should overwrite the default string with your string containg letters I, F, M and G in the order corresponding to pediree file.

PHENOTYPE_DATA_FILE (1 argument) specifies a file with quantitative records for individuals. Such quantitattive data can add extra information on genotype at a marker locus for an individual without genotype record. It is assumed that the trait is directly influnced by known diallelic marker. For the marker influencing the trait you should use Y_FUNCTIONAL option within MARKER_SPEC.

POLYGENIC_EFFECTS_MODE (1 argument) command specifies how polygenic additive effects are updated during calculations. The following options are allowed: *sample* or *random* (default) - a current value is replaced by a random sample, *expect* - a current value is replaced by an expected value calculated according to a particular model and current values for other model parameters, *freeze* or *frozen* - all polygenic values are equal to zero and kept unchanged. The last option is useful to remove polygenic additive effect from a model.

POSITION (2 arguments) is a subcommand used within the MARKER_SPEC command. See the MARKER_SPEC command for details.

POSITION_IN_FILE (2 arguments) is a subcommand used within the MARKER_SPEC command. See the MARKER_SPEC command for details.

PRINT_ALLELE_FREQ is a subcommand of the MARKER_SPEC command. See the MARKER_SPEC command for details.

PRINT_FIX_EFFECTS forces citius to print values of 'fixed' effects. The output is sent to a file named 'out.fix'. A row contains results from a particular iteration. A row starts with a value of general mean (or means for multivariate model). Then a row includes values for all but first level of second fixed effect. Again, for a model of n traits, there are n values for each level. Then goes the third fix effects, and so on. The first level of each fixed effect (except general mean) is zero. All fixed effects are automatically renumbered. The numbers are given in 'out.log' file.

PRINT_GENE_EFFECTS forces citius to print values of gene effects to a file named 'out.qtm'.

PRINT_POLYGENIC_ESTIMATES forces citius to print average vector of polygenic additive effects for all individuals. The averages are calculated from the values collected during iterations (except those produced in burn-in period). The output is sent to a file 'out.padd'.

PRINT_POLYGENIC_VARIANCE prints vales of the lower triangle of a polygenic additive (co)variance matrix. The output is sent to a file named 'out.va'.

PRINT_RESIDUAL_VARIANCE prints values of the lower triangle of a residual (co)variance

matrix. The output is sent to a file named 'out.ve'.

RESIDUAL_VARIANCE command serves to input residual (co)variance matrix for quantitative traits. The number of traits is decided by using the command PHENOTYPE_DATA_INPUTS. Only lower trangle of a matrix should by given. The matrix can be used as an initial value or kept unchanged (known variance) during calculations (see RESIDUAL_VARIANCE_MODE).

SSAMPLER_USE specifies the proportion of the use of the S-sampler (single site sampler). The command is used in a combination with the Lsampler_use, Msampler_use, A2_use and A3_use commands.

THINNING_VALUE specifies t parameter. Samples are collected every t-th iteration. Setting t=1 (default) means that all valid samples are used for estimation or printed to files. If the t-th sample has not been accepted in Metropolis-Hastings step, the sample is skipped.

USE_MATERNAL_ADDITIVE_EFFECT (no argument) inludes maternal additive effect in the model for all quantitative phenotypic traits. The initial value for maternal additive (co)variance matrix and covariance between direct and additive efects should be provided by using ADDITIVE_VARIANCE command.

Y_FUNCTIONAL is a subcommand of the MARKER_SPEC command. See the MARKER_SPEC command for details.

Genotype probabilities

Citius can be used to calculate genotype probabilities for each individual in a pedigree. Exact genotype probabilities are marginal probabilities obtained by summing over all possible genotype configurations which are consistent with data, weighting the configurations by their probability of occurence as calculated under the assumed genetic model. There are two methods possible: MCMC and the iterative peeling (IP) approach. MCMC method can be used for large but not huge pedigrees. For huge complex pedigrees the MCMC analysis may be impractical due the amount of time needed. If a pedigree is huge but its structure is simple the MCMC may still be practical. The calculation of genotype probabilities by the use of the MCMC method can be performed for both fixed and not fixed allele frequency. The iterative peeling method can be used to calculate genotype probabilities for a pedigree of any size. For huge complex pedigrees, as often analyzed in animal genetics, the iterative peeling approach may be the only possible option. The two approaches are compared in Table below.

MCMC analysis. This method can be used for large but not huge pedigrees. For huge complex pedigrees the MCMC analysis may be impractical due the amount of time needed. If a pedigree is huge but its structure is simple (there is no or only a few loops in the pedigree) the MCMC mathod may still be possible. The calculation of genotype probabilities by the use of the MCMC method can be performed for both fixed and not fixed allele frequency. Multilocus probabilities can be calculated. To indicate which locus should be considered for probability estimation you can use the G_PROB command within the MARKER_SPEC command. If the G_PROB option appears more than once, multilocus genotypes are considered. If genotype probabities are important for a part of a pedigree rather than for the entire pedigree you can provide a list of 'affecteds', for which the probabilities should be calculated. Use the AFFECTED_INDIVIDUAL_FILE command to present the list to citius. Note, citius uses the allele set-recoding by default to speed up calculations. The estimates, however, are always calculated for original alleles rather than set-recoded alleles. Therefore, there is usually no need to use the NO_SET_RECODING option (this option can be useful for IP algorithm).

Table. Comparison of two approaches for calculation of genotype probabilities.

	MCMC	**IP**
Number of loci	single or multilocus genotypes	single locus genotypes
Pedigree	impractical for huge complex pedigrees	any size and complexity
Allele frequency and gene effects	sampled based on sampled genotypes	updated by expectation calculated from current genotype probabilities
Precision	theoretically exact	approximate for looped pedigrees
Command	G_PROB	ITERATIVE_PEELING
Output file	out.gp	out.ip

Example. An example of a control file for MCMC estimation of genotype probabilities. The option G_PROB within the MARKER_SPEC command is used twice for D8S351 and D8S261 marker, therefore bilocus genotypes are considered. The genotype probabilities at the linked marker D8S552 are not considered but the marker contributes to genotype probability calculation at the flanking markers. Estimates are produced only for individuals listed in the file 'e1.aff'. The estimates are based on 30,000 samples, collecting every 15th sample after the burn-in period of 50,000 samples. Four samplers are used (L, M, A2 and A3) with the proportion of 10:3:1:1. The allele frequency are unknown and are sampled as well (default). Equal allele frequency are used as starting values (default).

```
pedigree_file   e1.ped
genotype_data_file   e1.dat
affected_individuals_file   e1.aff

burn_in 50000
thinning_value 15
number_of_samples 30000
display_value 100

number_of_markers 3
marker_spec name D8S552  position  19.5  28.4
marker_spec name D8S351  position  10.4  16.8  g_prob
marker_spec name D8S261  position  25.6  37.6  g_prob

Lsampler_use 10
Msampler_use 3
A2_use 1
A3_use 1
```

Example. Partial 'out.gp' file - an example of MCMC estimates of genotype probabilities. For the individual 666 the complete result is shown. For this individual only 4 bilocus genotypes were sampled - the highest probability (0.3434) was calculated for the genotype 123-2/124-5, where 123 and 2 are the paternal genes and 124 and 5 are the maternal genes at the two loci considered. For the individual E4 the number of different sampled genotypes was 134. The genotypes are sorted and the most probable genotypes appear first.

```
666 4
 123 124  2 5  0.3434
 124 123  2 5  0.3278
 123 124  5 2  0.1682
 124 123  5 2  0.1606
E4 134
 124 123  2 3  0.0580
 123 124  2 3  0.0556
...
```

The default behaviour of citius is that allele frequency is sampled (updated) at each iteration. The default initial allele frequency is $1/N_{alleles}$ for each allele. If the frequency for an allele is known you can use the known frequency as the initial value and keep the value constant during all iterations. To set the allele frequency to the known value, you can use the FIX_ALLELE_FREQ command (within the MARKER_SPEC command).

The results are sent to 'out.gp' file. The following is printed for each individual considered: ID of the individual, the number of different genotypes sampled, and then all sampled genotypes with the corresponding probabilities. The multilocus genotype is printed using the convention: paternal allele at first locus, maternal allele at first locus, paternal allele at second locus, maternal allele at second locus, etc. The markers are ordered according to their positions, from the smallest cM value to the highest cM value. You can find the order in 'out.log' file.

Iterative peeling (IP). The iterative peeling method can be used to calculate genotype probabilities at a single locus for a pedigree of any size and complexity. For huge complex pedigrees, as often analyzed in animal genetics, the iterative peeling approach may be the only practical method. For a zero-loop pedigree the IP converges to exact solutions. For a looped pedigree the IP calculates approximate solutions.

An initial allele frequency is $1/N_{alleles}$ for each allele. You may want to fix frequency of some alleles at some known values. For this use the FIX_ALLELE_FREQ command (within the MARKER_SPEC command).

The output from the IP calculation is sent to 'out.ip' file. The following pieces of information are printed: ID, marker name, the number of non-zero probability genotypes, and then all the genotypes with the calculated probabilities. For a genotype the paternal allele is printed first and the maternal is printed second.

Example. An example of a control file for the estimation of genotype probabilities by the use of the iterative peeling (IP) approach. The option ITERATIVE_PEELING within the MARKER_SPEC command is used for all three markers, therefore three separate calculations, for each locus in turn, are performed. Markers are always considered independent during IP calculation so the position of the markers are irrelevent. For the marker D8S261 the frequency of allele '5' (in the default genetic group) was set to 0.3, therefore, the frequencies of all other alleles, including the extra allele representing all not observed alleles, are estimated. The extra allele is not considered for the marker D8S551. The NO_SET_RECODING option is used for all markers to obtain results defined on original alleles.

```
pedigree_file                e1.ped
genotype_data_file           e1.dat

number_of_markers 3

marker_spec
    name D8S552
    no_set_recoding
    iterative_peeling
marker_spec
    name D8S551
    iterative_peeling
    no_set_recoding
    no_extra_allele
marker_spec
    name D8S261
    iterative_peeling
    fix_allele_freq 1
      default 5 0.3
```

Example. Partial 'out.ip' file - an example of genotype probability estimates calculated by the iterative peeling algorithm. The paternal allele is printed on the left and the maternal allele is printed on the right. For the individual AD there were 9 feasible genotypes at the D8S552 locus. The genotypes 11/11, 11/12, 12/11 and 12/12 are the most probable. The four genotypes have equal probaility of 0.138889. The extra allele, representing all unobserved alleles, was also considered and its code is '0'. Note, the genotypes are not sorted according to the probability.

```
AD  D8S552  9
11  11   0.138889
12  11   0.138889
0  11   0.097222
11  12   0.138889
12  12   0.138889
0  12   0.097222
11  0   0.097222
12  0   0.097222
0  0   0.055556
EV  D8S552  9
11  11   0.138889
12  11   0.138889
0  11   0.097222
11  12   0.138889
12  12   0.138889
0  12   0.097222
11  0   0.097222
12  0   0.097222
0  0   0.055556
...
```

Haplotype probabilities

Citius does not calculate haplotype probabilities directly. Instead you should calculate genotype probabilities and then use PERL utylity haplo.pl to calculate haplotype probabilities from estimated genotype probabilities. The following command

./haplo.pl out.gp 3

will calculate haplotype probabilities based on the results of genotype probabilities stored in 'out.gp' file. It is assumed that the probabilities were calculated for 3-locus joint genotypes.

IBD analysis

Citius can be used for MCMC estimation of the number of genes shared identical by descent (IBD) between relatives. One or many pairs of relatives can be considered and the IBD sharing statistics can be calculated at one or multiple points on a chromosome.

Points. You can use one of the two commands: IBD_POINTS or IBD_WALK. The IBD_POINTS command is more general and it can be used to analyze IBD sharing at one or multiple points on a chromosome. If you analyze IBD sharing at many evenly spaced points the IBD_WALK command is more convinient.

The IBD_POINTS command specifies one or many points on a chromosome. For these points IBD sharing statistics will be estimated. The IBD_POINT command is followed by

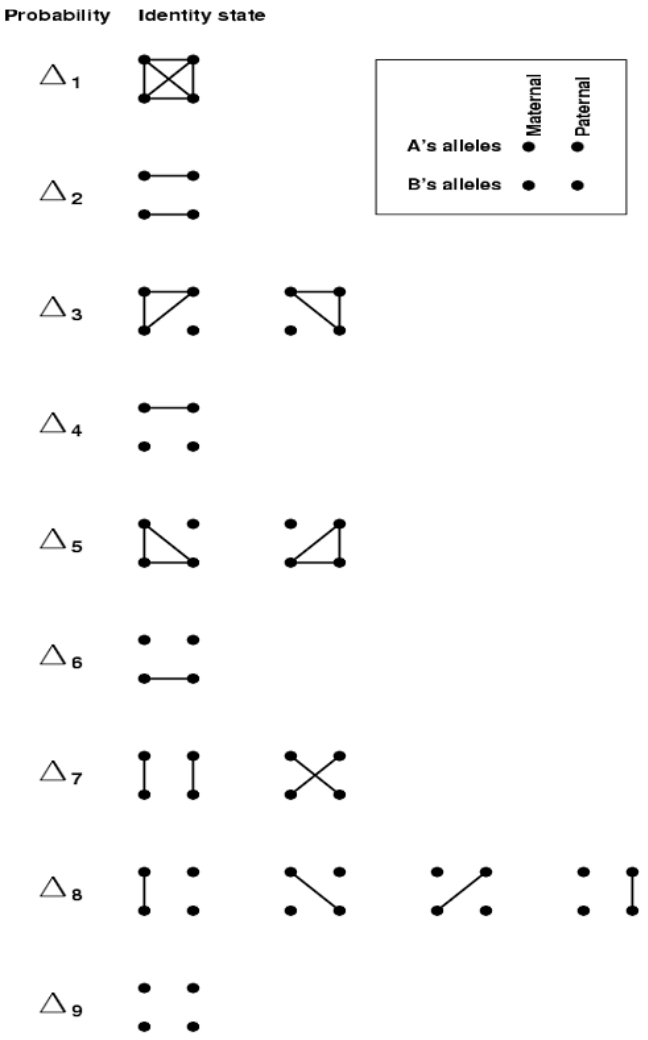


Figure. The 15 possible states of identity by descent for a locus in individuals A and B. Genes that are identical by descent are connected by lines.

the required number of points and positions on male and female linkage map (in cM). For example, the following command

IBD_POINTS 3 5 8 12 20 17 25

means that IBD sharing will be calculated at three points. On the male map the points are 5, 12 and 17 cM, and on the female map these are 8, 20 and 25 cM. Certainly, a point on the male and female map should be flanked by the same pair of markers.

If many points are considered, the IBD_WALK command may be more convinient. It specifies the starting point, the step, and the final point on a chromosome. For example, the following command

IBD_WALK 5 3 17

means that the points to be analyzed are: 5, 8, 11, 14 and 17. If a long fragment of a chromosome is analyzed there is possibility that some points will be flanked by different pairs of markers on male and female maps. This command is rather designed for sex average map. See the description of the MARKER_SPEC command to find out how to use sex average map.

Pairs. Many pairs of relatives can be considered. Be default citius analyzes all possible pairs, for example for a pedigree of N individuals $(N^2+N)/2$ pairs are considered. This number includes also N pairs for the relation between an individul with itself. Note, all pairs are taken into account ignoring the fact that many of them consist of unrelated

Example. The complete control file for possible IBD analysis. Three points are considered, on the male map they are: 19.0, 19.5 and 20.0, and on the female map the points are 27.9, 28.4 and 28.9. A list of 'affected' individuas is given in the 'e1.aff' file, therefore, only pairs formed by two 'affected' individuals are analyzed.

```
ibd_points 3 19.0 27.9  19.5 28.4  20.0 28.9

pedigree_file                e1.ped
genotype_data_file           e1.dat
affected_individuals_file    e1.aff

burn_in 50000
thinning_value 10
number_of_samples 3000
display_value 100

number_of_markers 3
marker_spec name D8S552  position  19.5  28.4
marker_spec name D8S551  position  10.4  16.8
marker_spec name D8S261  position  25.6  37.6

Lsampler_use 1
Msampler_use 3
A2_use 2    A3_use 2
```

individuals. Considering all possible pairs for a large pedigree is impractical and often unnecessery. To limit the number of pairs you can provide a list of 'affecteds'. If such a list is provided, only the pairs of affected individuals are considered. For example, if within N individuals there are N_A affected individuals, the number of pairs is $(N_A^2+N_A)/2$. The list, being a simple text file, can be presented to citius by the use of the AFFECTED_INDIVIDUALS_FILE command.

IBD probabilities. If a joint sample of segregation indicators at a locus is obtained, the founder genes are dropped down the pedigree. The number of founder genes is twice the number of founders. Sharing of founder genes is observed in pairs of relatives. The random state of founder genes in two individuals is classified into one of nine possible classes (see Figure). For a given pair the probability of each class is estimated as a ratio of number of samples classified to given class to the total number of samples. The coefficient of coancestry between individuals A and B due to segregation at the locus (point) under study is calculated as:

$$\Theta_{AB} = \Delta_1 + \tfrac{1}{2}(\Delta_3+\Delta_5+\Delta_7) + \tfrac{1}{4}\Delta_8$$

where Δs are the calculated class probabilities.

Output file. Results are sent to file 'out.ibd'. There are two possible forms of output. The standard and default output is for individual-individual IBD relationship and other is for gamete-gamete IBD relationship. The first format starts with two individuals' identification codes forming a pair, followed by the position of locus on male and female maps. The estimated coefficient of coancestry Θ_{AB} appears in the fifth column. Next, number of non-zero class probabilities is given, and for each non-zero probability class, the class number and the class probability appear.

Example. A partial output file from IBD analysis. Columns 1-2: pair of individuals. Columns 3-4: points (in cM) on male and female map. Column 5: the coefficient of coancestry. Column 6: the number of non-zero class probabilities. Next columns: non-zero class number and class probability. The IBD results are shown for two pairs: the first pair is EV and 666, and the other is for relationship of 666 with itself. Three points were considered. The points on the male maps are: 19.0, 19.5 and 20.0, an on the female map the corresponding points are 27.9, 28.4 and 28.9. For the first pair the coefficient of coancestry at the middle point is 0.109. For this pair, there were 5 non-zero class probabilities and the classes are 5, 6, 7, 8, and 9. The estimated probabilities for classes 1-4 are all zeros. For the relationship of 666 with itself, the coefficient of coencestry at the middle point is 0.530. Two non-zero class probabilities were calculated: for class 1 the probability is Δ_1=0.060 and for class 7 the probability is Δ_7=0.940.

```
...
EV 666    19.0  27.9  0.110  5  5 0.030  6 0.0340  7 0.017  8 0.349  9 0.570
EV 666    19.5  28.4  0.109  5  5 0.027  6 0.0330  7 0.017  8 0.350  9 0.573
EV 666    20.0  28.9  0.109  5  5 0.024  6 0.0320  7 0.020  8 0.348  9 0.576
...
666 666    19.0  27.9  0.532   2  1 0.064  7 0.936
666 666    19.5  28.4  0.530   2  1 0.060  7 0.940
666 666    20.0  28.9  0.528   2  1 0.056  7 0.944
```

If you need IBD relationship between gametes, please use the IBD_GAMETIC_MATRIX option in control file. Then the format for output is the following: (column 1) position on male map, (column 2) position on female map, (column 3) code of an individual being owner of first gamete, (column 4) code of an individual being owner of second gamete, (column 5) zero or one if first gamete is paternal or maternal, (column 6) zero or one if second gamete is paternal or maternal, (column 7) first gamete number, (column 8) second gamete number, (column 9) IBD probability for the two gametes. The gametes are numbered from 1 to 2N, the paternal and maternal gamete of first individual have numbers 1 and 2 and the gametes for the last individual have numbers 2N-1 and 2N.

Z-statistics. Citius can be hepful in gene mapping in complex diseases. The method of analysis is a version of the Affected Relative Pair (ARP) analysis. In this method a single statistics (Z) reflecting IBD sharing in pairs of affected relatives is calculated at given points on a chromosome. The points are defined by the IBD_POINTS or IBD_WALK command. The Z-statistics is calculated as follows: for a random sample of founder gene configuration in a pedigree dropped for a single point on a chromosome, for a pair of affected individuals i and j, the Z_{ij} value is calculated as

$$Z_{ij}=\sum_{a=1,2}\sum_{b=1,2}0.25\times I_{ia,jb}$$

where $I_{ia,jb}=1$ if the two genes (a and b) are IBD or 0 in other case. All Z_{ij} statistics are summarized (excluding parent-child pairs) as $Z=\sum Z_{ij}$. The final Z statistics at a point is calculated as an average from all samples. The Z statistics for all considered points are printed to 'out.arp' file. The fields in the file are: (1) the point (cM) on the male linkage map, (2) the point on the female map, and (3) the estimated Z statistics. The approximate distribution of the Z statistics can be calculated in separate runs after removing genotype records for affecteds.

Analysis of quantitative traits

Quantitative traits can be used in various ways:

- A quantitative trait with direct or indirect association with a studied molecular polymorphism (marker) can improve calculation of genotype probabilities. In consequence, final results for genotype probabilities or genes IBD can be improved.
- Quantitative traits can be tested for association with studied molecular polymorphism. Effect of a gene on phenotypic values can be estimated.
- Segregation analysis can be performed to test whether a trait (or set of traits) is shaped by an hypothetical unobserved major gene. Frequency of the gene and the gene effect can be estimated.
- Quantitative traits can be analysed under multitrait pure polygenic BLUP-AM model. Breeding values can be calculated under known variance components. Full bayesian approach can be used to estimate (co)variance components as well.

If a phenotype records are made available (see PHENOTYPE_DATA_FILE), the number of traits and fixed effects in the model are determined by the command PHENOTYPE_DATA_INPUTS. Then a deafult model is a pure polygenic additive model. To include an effect of a gene, use the command MARKER_SPEC / Y_FUNCTIONAL. More than one gene can be labelled Y-FUNCTIONAL. If you know the effects of a gene, you can input them by the use of GENE_EFFECTS command. The command is also useful to initialize gene effects (deault values are zeros). To remove polygenic component from a model, use POLYGENIC_EFFECT_MODE FREEZE. You can initialize variance components through POLYGENIC_VARIANCE and RESIDUAL_VARIANCE commands.

The regular behaviour of citius is that it does not print values of parameters to a file. You can force citius to print values of various model parameter by using different PRINT_... commands.

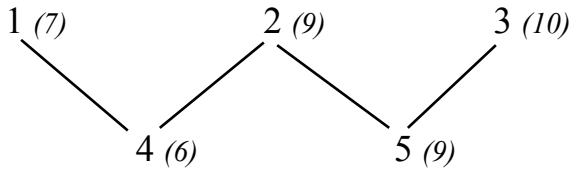
Be default, citius performs full bayesian analysis. Different model parameters are updated by random values. You can change this behaviour by the use of different ..._MODE commands. You have full control on what citius does. You can replace some of parameter values by expected values rather than random values and keep other values fixed through all iterations. Please be advised that not all scenarios lead to correct algorithm.

Relevent commands

- PHENOTYPE_DATA_FILE
- PHENOTYPE_DATA_INPUTS
- MISSING_Y_CODE
- ADDITIVE_VARIANCE

- POLYGENIC_VARIANCE_MODE
- PRINT_POLYGENIC_VARIANCE
- PRINT_POLYGENIC_ESTIMATES
- RESIDUAL_VARIANCE
- RESIDUAL_VARIANCE_MODE
- PRINT _RESIDUAL_VARIANCE
- MARKER_SPEC / Y_FUNCTIONAL
- MARKER_SPEC / GENE_EFFECTS
- GENE_EFFECTS_MODE
- PRINT_GENE_EFFECTS
- USE_MATERNAL_ADDITIVE_EFFECT

Example. Unitrait BLUP-AM model. Consider a simple example from book *Genetics and analysis of quantitative traits* by Lynch and Walsh. There are 5 individuals and each individual has a single measurment (in *italic*) and the only fixed factor is the mean. In the original example the ratio of residual and polygenic variance was equal to 1, here two arbitrary values were used for variance components.



Control file

```
pedigree_file  5.ped
phenotype_data_file  5.dat
phenotype_data_inputs  IY

fix_effect_mode  expect
polygenic_effect_mode  expect
polygenic_variance_mode  freeze
residual_variance_mode  freeze

print_polygenic_estimates

burn_in 1000
thinning_value 1
number_of_samples 1
display_value 10
polygenic_loop_iterations 1

residual_variance
0.5

additive_variance
0.5
```

Pedigree file

```
4 1 2
5 3 2
```

Data file

```
1    7.0
2    9.0
3   10.0
4    6.0
5    9.0
```

Solutions: file 'out.padd'

```
1    -0.960813
2    0.075472
3    0.885341
4    -1.062409
5    0.552975
```

Example. Two-trait BLUP-AM. The example is from the book 'Linear models for the prediction of animal breeding values', 2nd edition, by R.A. Mrode. WWG=pre-weaning gain, PWG=post-weaning gain.

Calf	Sex	Sire	Dam	WWG	PWG
4	Male	1	-	4.5	6.8
5	Female	3	2	2.9	5.0
6	Female	1	2	3.9	6.8
7	Male	4	5	3.5	6.0
8	Male	3	6	5.0	7.5

G = 20 18 R = 40 11
18 40 11 30

Control file

```
pedigree_file                86.ped
phenotype_data_file          86.dat
phenotype_data_inputs        IFYY

fix_effect_mode         expect
polygenic_effect_mode   expect
polygenic_variance_mode freeze
residual_variance_mode  freeze

print_fix_effects
print_polygenic_estimates

burn_in 500
thinning_value 1
number_of_samples 1
display_value 10
polygenic_loop_iterations 1

residual_variance
40.0
11.0 30.0


additive_variance
20.0
18.0 40.0
```

Pedigree file

```
4 1 0
5 3 2
6 1 2
7 4 5
8 3 6
```

Data file

```
4  M  4.5  6.8
5  F   2.9  5.0
6  F   3.9  6.8
7  M  3.5  6.0
8  M  5.0  7.5
```

Solutions: file 'out.padd'

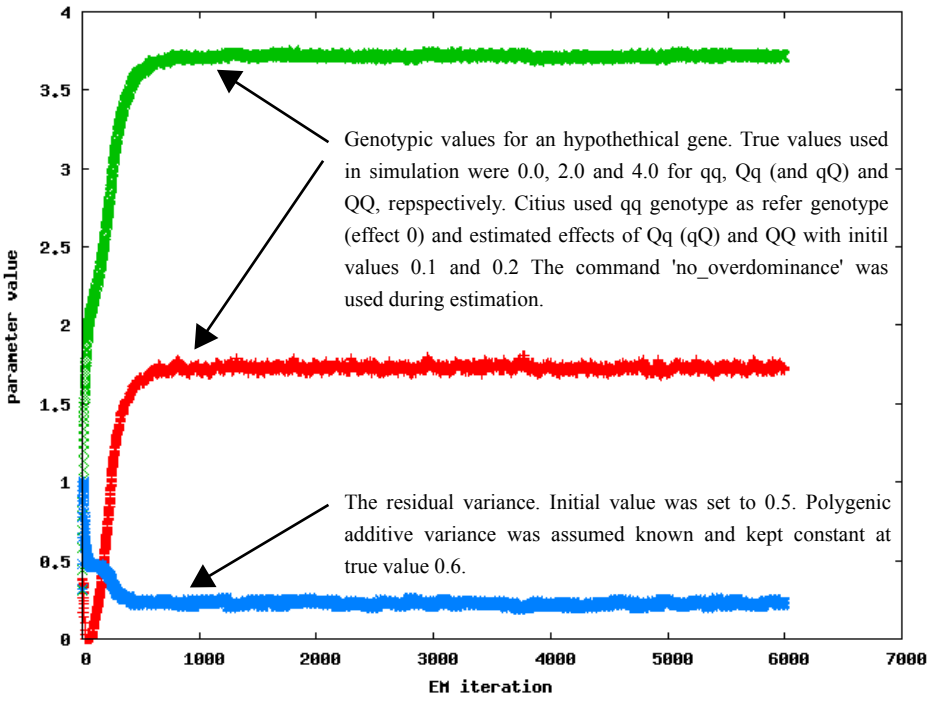
```
1   0.150916 0.279598
3   -0.078392 -0.170341
2   -0.015393 -0.007610
4   -0.010239 -0.012671
5   -0.270331 -0.477830
6   0.275808 0.517238
7   -0.316118 -0.478984
8   0.243756 0.391962
```

Example. Monte Carlo estimation of mixed model that includes major gene effect and polygenic effect. Guo and Thompson (1994, Biometrics 50: 417-432) presented a Monte Carlo method, using jointly the EM algorithm and Gibbs sampler, for estimation of mixed models. The algorithm of Guo and Thompson is shortly presented. To use the algorithm the control file contains the *increment* option and THE NUMBER_OF_INCREMENTS command. At each EM iteration, 400 Gibbs samples were drawn (N=400). To perform the Monte Carlo EM for 6000 iterations, we generated total 2,4 mln samples. Guo and Thompson used single site sampler for genotypes while citius sample genotypes for entire pedigree jointly (better mixing). Therefore, there is no need to have additional cycles between samples of **G**. For polygenic effects, however, citius uses single site sampler that needs some spacing, and we use S=10. We printed every new EM update (THINNING_VALUE=400).

...
4. Set initial parameter estimates, p, m_i, v_a, v_e
...
7. Next EM iteration step
  Set $p^*=m_1^*=m_2^*=m_3^*=m_4^*=v_a^*=v_e^*=0$
    For n=1 to N (Monte Carlo sample size)
      Gibbs sample (**G**, **a**) ;
        for l=1 to S (the chosen spacing)
          update genotypes
          update polygenic additive effects
        next l
      After *c*th cycle, we have configuration (**G**, **a**)
      increment p* by E(p|**G**)/N
      increment m_i* by E(m_i| **y, G**, **a**, θ)/N
      increment v_a* by E(v_a| **a**)/N
      increment v_e* by E(v_e| **y, G**, **a**, θ)/N
    next n
    update parameter estimates p=p*, $m_i=m_i$*, $v_a=v_a$*, $v_e=v_e$*
  Go to step 7

p = gene frequency
m_i = contribution of genotype i to the phenotype (i=1,2,3,4)
v_a = polygenic additive variance
v_e = residual variance
G = configuration of major genotypes
a = vector of additive polygenic effects
y = vector of measurments
θ = (p, m_i, v_a, v_e)

```
pedigree_file                230-member.ped
phenotype_data_file          simdat3
phenotype_data_inputs        IY

fix_effect_mode         increment
gene_effect_mode        increment
polygenic_variance_mode freeze
residual_variance_mode  increment

print_gene_effects
print_residual_variance

burn_in 0
thinning_value 400
number_of_samples 2400000
display_value 50
number_of_increments 400
polygenic_loop_iterations 10

number_of_markers 1
marker_spec
  name SomeGene
  position  0.0  0.0
  positions_in_file 0 0
  y_functional
  no_extra_allele
  no_set_recoding
  print_allele_freq
  allele_freq_mode increment
  initial_gene_effects   .1  .1  .2
  no_overdominance
  no_imprinting

polygenic_variance 0.6
residual_variance  0.5
```


parameter value
EM iteration
0
0.5
1
1.5
2
2.5
3
3.5
4
0
1000
2000
3000
4000
5000
6000
7000
Genotypic values for an hypothethical gene. True values used in simulation were 0.0, 2.0 and 4.0 for qq, Qq (and qQ) and QQ, repspectively. Citius used qq genotype as refer genotype (effect 0) and estimated effects of Qq (qQ) and QQ with initil values 0.1 and 0.2 The command 'no_overdominance' was used during estimation.
The residual variance. Initial value was set to 0.5. Polygenic additive variance was assumed known and kept constant at true value 0.6.

References

Fernandez SA, Fernando RL, Guldbrandtsen B, Totir LR, Carrquiry AL (2001): Sampling genotypes in large pedigrees with loops. Genet Sel Evol, 33: 337-367.

Lange K, Goradia TM (1987): An algorithm for automatic genotype elimination. Am. J. Hum. Genet. 40: 250-256.

Matsumoto M and Nishimura T (1998): Mersenne Twister: A 623-Dimensionally Equidistributed Uniform Pseudo-Random Number Generator. ACM Transactions on Modeling and Computer Simulation, Vol. 8, No. 1, January 1998, pp 3--30.

O'Connell J.R, Weeks DE (1995): The VITESSE algorithm for rapid exact multilocus linkage analysis via genotype set-recoding and fuzzy inheritance. Nature Genetics 11: 4002-408.

Szydlowski M, Gengler N (2008): Sampling genotype configurations in a large complex pedigree. Journal of Animal Breeding and Genetics (accepted).

Thomas A, Gutin A, Abkevich V, Bansal A (2000): Multilocus linkage analysis by blocked Gibbs sampling. Statistics and Cmputing 10: 259-269.

Thompson EA, Heath SC (1999): Estimation of conditional multilocus gene identity among relatives. Statistics in Molecular Biology, IMS Lecture Notes - Monograph Series 33: 95-133.

Tierney L (1994): Markov chains for exploring posterior distributions. Ann Statist 22: 1701-1762.

van Arendonk JAM, Smith C, Kennedy BW (1989): Method to estimate genotype probabilities at individual loci in farm livestock. Theor Appl Genet 78: 735-740.